

# STAT 401 Chapter 6 except 6.5

Zepu Zhang

October 25, 2010

In this chapter we tidy up what we learned in Chapter 2, but in matrix notation. The book has a tendency to fall back on element-wise summations. We should abstain from using the  $\sum$  sign. Appreciate the elegance made possible by matrix notation. Most of the formulas are identical to those in lecture note ch05b.

## 1 Model formulation

We use  $p \geq 2$  predictors (the first is the constant 1) to model a response variable  $Y$ . The predictor vector can be written  $\vec{x} = [1, x_1, \dots, x_{p-1}]'$ . There are  $p$  coefficients (one for each predictor component), written as vector  $\vec{\beta} = [\beta_0, \beta_1, \dots, \beta_{p-1}]'$ . The statistical model for  $Y$  is

$$Y = \vec{x}'\vec{\beta} + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2).$$

To express the model for a data set  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , define vectors  $\vec{Y} = [Y_1, \dots, Y_n]'$ ,  $\vec{\epsilon} = [\epsilon_1, \dots, \epsilon_n]'$ , and “design matrix”

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix},$$

where  $[1, x_{i,1}, \dots, x_{i,p-1}]'$  (i.e. the  $i$ th row) is the predictor vector corresponding to  $Y_i$ . Then the model is written

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}, \quad \text{where } \vec{\epsilon} \sim N(\vec{0}, \sigma^2 \mathbf{I})$$

or

$$\vec{Y} \sim N(\mathbf{X}\vec{\beta}, \sigma^2 \mathbf{I}).$$

### Generality of this formulation

(page 218–221)

The “predictors” need not be distinct physical quantities.

- They can be different forms of one quantity, e.g. elevation, elevation<sup>2</sup>, log elevation.

- They can be interactions, e.g. weight  $\times$  age.
- They can be qualitative, e.g. color (yellow/red/green), gender (M/F).

A qualitative predictor needs to be coded as numbers in some way. See page 218–219. Basic idea: introduce  $k - 1$   $X$  components to code a qualitative variable that has  $k$  possible values.

- $Y$ , like  $X$ , may also be transformations of the more “natural” quantity, e.g.  $Y = \log(\text{earthquake energy})$ .

The defining characteristic of the model is that it is linear in the model parameters,  $\vec{\beta}$ .

Also note that the predictors are not random variables; instead, they are just numbers. Therefore it is conceptually no problem if two predictors are values based on related things, for example one predictor is “height” and another is “height<sup>2</sup>”. However, using such closely related predictors may be inefficient.

## 2 Least squares estimation

### 2.1 LS estimation of model parameter

From the normal equation (which is equivalent to the LS criterion)

$$\mathbf{X}'\mathbf{X}\vec{\beta} = \mathbf{X}'\vec{Y}$$

we get the estimator

$$\vec{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\vec{Y}.$$

(Formula (6.25) on page 223 is wrong.)

(You should be able to derive all the formulas after this point in this note. Don’t take them as something opaque.)

### 2.2 Sampling distribution of parameter estimators

Because  $\vec{Y} \sim N(\mathbf{X}\vec{\beta}, \sigma^2\mathbf{I})$ , we see

$$\vec{\beta} \sim N(\vec{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

## 2.3 Fitted values

$$\vec{\hat{Y}} = \mathbf{X}\vec{\hat{\beta}} = \mathbf{H}\vec{Y}, \quad \text{where } \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Exercise Show  $\vec{\hat{Y}} \sim N(\vec{Y}, \sigma^2 \mathbf{H})$ .

The above is the vector of fitted values corresponding to the observations. The fitted values for a new  $X$  value, say  $\vec{x}_*$  (a column vector), is

$$\hat{Y}_* = \vec{x}_*' \vec{\hat{\beta}} = \vec{x}_*' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \vec{Y}$$

This is again a normal variable and we can find its sampling mean and variance (see below).

## 2.4 Residuals

$$\vec{e} = \vec{Y} - \vec{\hat{Y}} = (\mathbf{I} - \mathbf{H})\vec{Y}.$$

Exercise Show  $\mathbf{I} - \mathbf{H}$  is symmetric, and  $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{H}$ .

Exercise Show  $\vec{e} \sim N(\vec{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$ .

## 2.5 Estimation of $\sigma^2$

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-p} = \frac{\vec{e}'\vec{e}}{n-p} = \frac{1}{n-p} \vec{Y}'(\mathbf{I} - \mathbf{H})\vec{Y}.$$

This estimator is also called  $S^2$ .

- Note
1. We know  $E(S^2) = \sigma^2$ , i.e. this is an unbiased estimator.
  2. Plug in the actual observations  $\vec{y}$ , we can write the estimated value of  $\sigma^2$  (the “estimate”) as

$$s^2 = (n-p)^{-1} \vec{y}'(\mathbf{I} - \mathbf{H})\vec{y}.$$

3. In SLR, we had  $p = 2$  (two regression coefficients), hence we were using  $SSE/(n-2)$ .

## 3 Inferences

Recall that we refer to three things by ‘inference’: point estimation, confidence interval, and hypothesis tests. The basis of all three is a “good” (usually unbiased) point estimator and its sampling distribution. Once we have these, CI and tests follow the routine that we learned in Chapter 2.

### 3.1 Inferences about model parameters

Inferences (confidence interval, test) about an individual parameter ( $\beta_i$ ,  $i = 0, 1, \dots, p-1$ ) is based on its normal sampling distribution.

Note the estimators are all unbiased. Their variances are the diagonal elements of their cov matrix,  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ .

We can't write out analytically the diagonal elements because we can't write out the elements of an arbitrary inverse matrix. Don't worry. Just take the numbers from the calculated inverse matrix. Don't go back to the elementary formulas in Chapter 2 for  $\text{var}(\hat{\beta}_0)$ ,  $\text{var}(\hat{\beta}_1)$ , etc.

We need to substitute the estimate  $s^2$  for the unknown  $\sigma^2$ , then use a  $t$  distribution with  $n - p$  df.

### 3.2 Estimation of mean response

Our point estimator of  $E(Y)$  at  $\vec{x} = \vec{x}_* = [1, x_{1*}, \dots, x_{(p-1)*}]'$  is

$$\hat{Y}_* = \vec{x}_*' \vec{\hat{\beta}}.$$

This is a linear function of  $\vec{\hat{\beta}}$ , hence is a normal variable (before plugging in the actual estimates  $\vec{b}$  or observations  $\vec{y}$ ).

**Exercise** Check the right-hand side is 1 by 1 as needed.

Parameters for the distribution of  $\hat{Y}_*$  are

$$E(\hat{Y}_*) = \vec{x}_*' E(\vec{\hat{\beta}}) = \vec{x}_*' \vec{\beta}$$

(unbiasedness) and

$$\text{var}(\hat{Y}_*) = \vec{x}_*' \text{cov}(\vec{\hat{\beta}}) \vec{x}_* = \sigma^2 \vec{x}_*' (\mathbf{X}'\mathbf{X})^{-1} \vec{x}_* \quad (1)$$

After we substitute  $s^2$  for  $\sigma^2$ , we'll use a  $t$  distribution with  $n - p$  df to conduct inferences.

### 3.3 Prediction of new observations

Say we're going to make a new observation at  $\vec{x} = \vec{x}_*$ . The new observation is a particular value that happens to be taken by the random variable

$$Y_* = \vec{x}_*' \vec{\beta} + \epsilon_*.$$

It's natural that we use the following point “guess”:

$$\hat{Y}_* = \vec{x}'_* \vec{\hat{\beta}}.$$

Examine the difference  $Z = Y_* - \hat{Y}_* = \vec{x}'_* \vec{\beta} + \epsilon_* - \vec{x}'_* \vec{\hat{\beta}}$ . This is a normal random variable because it is const + normal – normal. Specifically,

$$E(Z) = \vec{x}'_* \vec{\beta} + E(\epsilon_*) - \vec{x}'_* E(\vec{\hat{\beta}}) = \vec{x}'_* \vec{\beta} + 0 - \vec{x}'_* \vec{\beta} = 0$$

$$\text{var}(Z) = \text{var}(\epsilon_*) + \text{var}(\vec{x}'_* \vec{\hat{\beta}}) = \sigma^2 + \sigma^2 \vec{x}'_* (\mathbf{X}' \mathbf{X})^{-1} \vec{x}_*.$$

In the above we have made use of the independence between  $\epsilon_*$  and  $\hat{Y}_*$ , and the result (1).

Based on these results we construct a “prediction interval” for  $Z$ , and that gives us a prediction interval for  $Y_*$ , since  $Z = Y_* - \hat{Y}_*$ . After substituting  $s^2$  for  $\sigma^2$ , a  $t$  distribution with  $n - p$  df (instead of a normal) is used.

## 4 Diagnostics and remedial measures

1. Scatter plot matrix.
2. Residual plots: against fitted  $Y$ 's, against each predictor  $X$ , against un-used variables (time, space, interaction, etc.). Watch out for curvature patterns, and constancy of variance.
3. Normality plot for residuals.
4.  $F$  test for lack of fit. (Later.)
5. Remedial measures: introduce additional predictors, transformations, etc.

## 5 Example

Section 6.9, page 236–247.